Structure and Dynamics of Diffusion Networks

We are interested in assaying and controlling diffusion processes in a broad range of domains, addressing two problems: network inference from diffusion traces [1, 2, 3] and spread maximization [5].

Network Inference Observing a diffusion process often consists of noting when nodes (people, blogs, etc.) reproduce a piece of information, get infected by a virus, or buy a product. Epidemiologists can observe when a person becomes ill but they cannot tell who infected her, and exactly when. We observe when a blog mentions a piece of information. However if, as is often the case, the blogger does not link to her source, we do not know where she acquired the information or how long it took her to post it. Finally, viral marketers can track when customers buy products or subscribe to services, but typically cannot observe who influenced customers and how long they took to make up their minds. We observe where and when but not how or why information propagates through a population of individuals. The hidden mechanism underlying the process is of outstanding interest, since understanding diffusion is necessary for stopping infections, predicting meme propagation, or maximizing sales of a product.

In our work, we formulated a generative probabilistic model of diffusion that aims to describe realistically how diffusion occurs over time in a network. First, we developed two algorithms, NETINF [1, 3] and MULTI-TREE [6], which use submodular optimization to infer the structure of a diffusion network from diffusion traces. We validated our algorithm by tracing information cascades in a set of 170 million blogs and news articles from 3.3 million sites over one year. The diffusion network of news tends to have a core-periphery structure with a small set of core media sites that diffuse information to the rest of the Web. These sites tend to have stable circles of influence with more general news media sites acting as connectors between them[1].

However, both algorithms force the transmission rates between all nodes to be fixed. To overcome this limitation, we then developed the algorithm NETRATE [2], which allows transmission at different rates across different edges so that we can infer temporally heterogeneous interactions within a network. NETRATE infers both structure and the temporal dynamics using convex optimization. Finally, many times networks over which propagation occurs change over time. For example, a blog can increase its popularity abruptly after one of its posts turns viral and this may create new edges in the underlying network. Therefore, we developed the algorithm INFOPATH, which uses stochastic convex optimization to be able to infer time-varying networks [4]. We study the evolution of information pathways in the Web. We find that information pathways for general recurrent topics are more stable across time than for ongoing news events. Clusters of news media sites and blogs often emerge and vanish in matter of days for ongoing news events. Major events involving civil population, such as the Libyan's civil war, lead to an increased amount of information pathways among blogs as well as in the overall increase in the network centrality of blogs and social media sites.

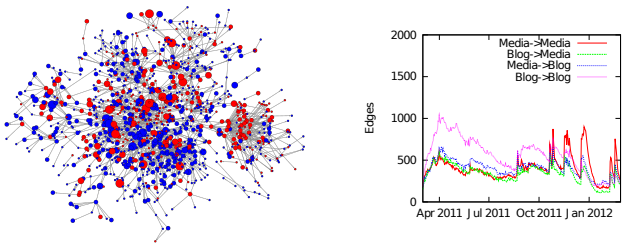

Figure 0.1: Left figure shows the network of 5,000 most active news sites (red) and blogs (blue) inferred by NETINF. Right figure shows the types of links for 5,000 most active sites for Syria's uprise inferred by INFOPATH.

Influence Maximization Spread maximization tackles the problem of selecting the most influential source node set of a given size in a diffusion network. A diffusion process that starts in such an influential set of nodes is expected to reach the greatest number of nodes in the network. Although the problem depends dramatically on the underlying temporal dynamics of the network, this still remains largely unexplored to date. In our work, we developed a method for influence maximization, INFLUMAX [5], which accounts for the temporal dynamics underlying diffusion processes [2]. Our method allows us to evaluate the influence of any set of source nodes in a network analytically and it finds the near-optimal set of nodes that maximizes influence by exploiting the submodularity of our objective function. Experiments on synthetic and real diffusion networks show that our algorithm outperforms other state-of-the-art algorithms by at least 20%.

[1] Part of our work won a best paper award honorable mention at KDD 2010

Publications

Journal Articles

[1] M Gomez Rodriguez, J Leskovec, and A Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data*, 5(4:21), 2 2012. 1

Articles in Conference Proceedings

[2] M Gomez Rodriguez, D Balduzzi, and B Schölkopf. Uncovering the temporal dynamics of diffusion networks. In L Getoor and T Scheffer, editors, *28th International Conference on Machine Learning (ICML)*, pages 561–568, Bellevue, WA, USA, 7 2011. International Machine Learning Society. 1

[3] M Gomez Rodriguez, J Leskovec, and A Krause. Inferring networks of diffusion and influence. In B Rao, B Krishnapuram, A Tomkins, and Q Yang, editors, *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 1019–1028, Washington, DC, USA, 7 2010. ACM Press. 1

[4] M Gomez Rodriguez, J Leskovec, and B Schölkopf. Structure and dynamics of information pathways in on-line media. In *6th ACM International Conference on Web Search and Data Mining (WSDM 2013)*, Rome, Italy, 2013. 1

[5] M Gomez Rodriguez and B Schölkopf. Influence maximization in continuous time diffusion networks. In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 313–320, Edinburgh, Scotland, 7 2012. Omnipress. 1

[6] M Gomez Rodriguez and B Schölkopf. Submodular inference of diffusion networks from multiple trees. In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 489–496, Edinburgh, UK, 7 2012. Omnipress. 1